# COPD 多维特征提取与集成诊断方法 *

房有丽 [1a, 1b, 2]，王　红 [1a, 1b, 2†]，狄瑞彤 [1a, 1b, 2]，王露潼 [1a, 1b, 2]，宋永强 [1a, 1b, 2]

(1. 山东师范大学 a. 信息科学与工程学院; b. 生命科学研究院，济南 250358; 2. 山东省分布式计算软件新技术重点实验室，济南 250358)

**摘　要：**慢性阻塞性肺疾病(COPD)是一种可导致患者呼吸功能逐渐下降的慢性肺部疾病，需要借助于大数据分析及算法帮助医生对疾病更加准确地进行诊断。目前对 COPD 的研究存在局限性，一方面，研究成果只利用数据分析单一特征对疾病的影响，另一方面研究成果仅通过简单算法模型对病例数据验证，因此提出了 COPD 多维特征提取与集成诊断方法。首先，提出最大依赖度 MDF-RS 算法，提取多维特征的最优组合；其次，提出 DSA-SVM 集成模型，构建分类器进行诊断及预测；最后，利用交叉验证方法验证准确率等各项性能指标。通过实验对比验证了该算法的有效性。

**关键词：**慢性阻塞性肺疾病；多维特征；集成方法；交叉验证

**中图分类号：**TP391　　doi: 10.3969/j.issn.1001-3695.2018.03.0204

## Multidimensional feature extraction and integrated diagnosis of COPD

Fang Youli[1a, 1b, 2], Wang Hong[1a, 1b, 2†], Di Ruitong[1a, 1b, 2], Wang Lutong[1a, 1b, 2], Song Yongqiang[1a, 1b, 2]

(1. *a. School of Information Science & Engineering, b. College of Life Science Shandong Normal University, Jinan* 250358, *China*; 2. *Shandong Provincial Key Laboratory of Distributed Computing Software, Jinan* 250358, *China*)

**Abstract:** Chronic obstructive pulmonary disease (COPD) is a chronic lung disease that can lead to a gradual decline in respiratory function. Therefore, big data analysis and algorithms are needed to help doctors diagnose diseases more accurately. At present, there are limitations to the study of COPD: On the one hand, the research results only use data to analyze the impact of single features on the disease; on the other hand, the research results are only verified by simple algorithm models for case data. Therefore, this paper proposes a COPD multi-dimensional feature extraction and integrated diagnosis method. First, the MDF-RS algorithm is proposed to extract the optimal combination of multi-dimensional features. Secondly, the DSA-SVM integrated model is proposed to construct the classifier for diagnosis and prediction. Finally, the cross-validation method is used to verify the accuracy and other performance indicators. The experimental comparison shows the effectiveness of the proposed algorithm.

**Key words:** COPD; multidimensional features; integration methods; cross validation

## 0　引言

慢性阻塞性肺疾病（COPD）是一种可导致患者呼吸功能逐渐下降的疾病，其已成为全球第四大致死疾病[1]，全球目前约有超过 1.7 亿 COPD 患者。COPD 的病情发展是渐进性的过程：早期，COPD 症状并不明显，主要是咳嗽、咳痰，患者不易察觉，是最佳治疗时机；中期，随着病情的加重，患者可能出现活动后呼吸困难，气道阻塞加重、肺组织弹性损坏，达到不可逆转阶段，各种药物都难以发挥作用；晚期，可出现肺心病、呼吸衰竭等并发症，治疗若不及时，会严重影响患者的生活质量和身心健康。

所以 COPD 的早期发现非常重要，需要长期稳定的管理患者病情。如果不预防不管理，随着疾病的进一步发展，特别是发生急性加重就会给患者带来更大的危害。急性加重是 COPD 患者的咳嗽、咳痰、呼吸困难、胸闷、喘息等症状在短期内急剧恶化，并可能导致治疗措施的改变。随着计算机数据挖掘技术的发展，该类问题成为计算机领域一个研究热点。

目前，数据挖掘技术已经广泛应用于对 COPD 病理分析及

**收稿日期：**2018-03-25；**修回日期：**2018-05-08　**基金项目：**国家自然科学基金资助项目（61672329，61373149，61472233，61572300，81273704）；山东省科技计划资助项目（2014GGX101026）；山东省教育科学规划资助项目（ZK1437B010）；山东省泰山学者基金资助项目（TSHW201502038，20110819）；山东省精品课程资助项目（2012BK294，2013BK399，2013BK402）

**作者简介：**房有丽（1991-），女，山东潍坊人，硕士研究生，主要研究方向为数据挖掘、机器学习（1521055775@qq.com）；王红（1996-），女教授，博导，博士，主要研究方向为大数据、复杂网络、数据挖掘等；狄瑞彤（1993-），女，硕士研究生，主要研究方向为数据挖掘；王露潼（1994-），女，硕士研究生，主要研究方向为数据挖掘、机器学习、复杂网络；宋永强（1994-），男，硕士研究生，主要研究方向为数据挖掘、机器学习、复杂网络.

临床诊断等研究领域[2]。主要在两个方面研究：a）利用现有的数据分析工具对电子病例数据分析，以挖掘单一特征对疾病的影响；b)通过简单模型验证 COPD 的患者预后风险效果。

本文的主要贡献包括：a）提出 MDA-RS 算法，提取 COPD 的最优特征子集，以支持更好的分类结果;b）提出 DSA-SVM 混合模型，对慢性阻塞性肺疾病进行分类和预测;c)进行大量实验，证明我们方法的有效性。

## 1　相关工作

近年来，通过对 COPD 数据分析及特征表现如何辅助医生诊断成为一个研究热点。研究者们主要在分析特征影响因素及不同阶段疾病风险预测方面做了大量的工作，并获得了较好效果。

Himes 等人[3]利用从哮喘病人的病历中提取的特征和人口统计学信息，建立了预测 COPD 的模型，并使用该模型预测独立哮喘患者的 COPD 预测准确性。 在这个模型中，年龄，性别，种族，吸烟史等 8 种特征预测了 COPD 的风险。通过多次实验，该模型准确率达到 0.83。

Hoogendoorn 等人[4]使用 COPD 数据源进行数据分析发现重要预测因素包括咳嗽和喘息，咳痰，步行 6 分钟，使用吸入性皮质类固醇和氧饱和度。预测结果符合真实病例情况，但此外，低体重指数，心血管疾病和肺气肿是二级保健患者住院治疗的重要预测因素。

郭慧敏等人[5]使用 R 语言做模型的识别、模型的参数估计与检验，以每月的入院人次构成时间序列，建立 ARIMA 模型对 COPD 的预测，结果显示 ARIMA 模型较好地拟合 COPD 入院人次并进行短期预测，模型显示了 2016 年该院的 COPD 的入院有所上升，为医院合理利用医疗资源提供了有力依据。

前面介绍了一部分研究者通过分析特征因素对疾病的影响。另外，还有其他工作者对 COPD 不同阶段的风险预测分析。例如文献[6~8]对于风险分层处理的先决条件进行分析，使 COPD 患者得到更好的诊断及治疗，以避免原有的风险可能导致更高的健康相关的生活质量及更长的寿命和更低的医疗成本。文献[9,10]通过预测模型对风险分层治疗进行实验验证及对比。

Mega[11]等人研究评估 BODE 指数（一种预测死亡率的多维分级系统）的能力，以预测 COPD 患者的病情状况。结论描述BODE 指数是 COPD 急性加重次数和严重程度的一个更好的预测指标。

通过总结前人的工作发现，研究者分别从患者数据分析及疾病风险的预测两个方面进行研究。本文提出了基于 DSA-SVM 算法的混合决策方法对慢性阻塞性肺疾病的诊断并构建分类器，通过属性最大依赖度 MDA-RS 算法对多维特征的提取，并用交叉验证方法验证了准确率等各项指标。

## 2　COPD 方法

本文的目标是通过对慢性阻塞性肺疾病患者数据分析，提取多维特征的特征子集，利用混合决策模型 DSA-SVM 算法对疾病的诊断预测。为了实现这个目标，有四个问题需要解决，步骤如下：数据预处理；利用 MDA-RS 算法对多维特征选择；优化参数算法 DSA；构建混合决策模型 DSA-SVM 分类器。

### 2.1　数据预处理

数据预处理的目的是为了提高数据质量，使数据挖掘的过程更加有效，更加容易，同时也提高挖掘结果的质量。数据预处理的对象主要是噪声数据、缺失数据。常用的数据预处理技术主要包括：数据清洗、相关分析和数据变换等。

本文中对原始数据做了适当的预处理，具体步骤如下所示：

a)将分类属性转换为数字数据项。我们用数值来表示每个分类值，例如，吸烟用 1 表示，不吸烟用 0 表示。

b)对原始的缺失数据通过临近值或者均值填充。例如 COPD 数据集第 8、16 个特征分别具有 37、23 个缺失值，可以利用该属性的众数填充。

c)数据归一化[12]。例如对第一秒用力呼气容积与用力呼气容量比值（FEV1/FVC），可以归一到（0~1）内数据通过归一化处理有利于计算，并提高计算精度精度。其中数据归一化的公式如式（1）所示，$X_{norm}$ 为归一化后的数据，$X$ 为原始数据，$X_{\max}$，$X_{\min}$ 分别为原始数据的最大值和最小值。

$$X_{norm}=\frac{X-X_{\min}}{X_{\max}-X_{\min}} \tag{1}$$

### 2.2　多维特征选择

图像处理、信息检索以及生物信息学等技术的发展,产生了以超大规模特征为特点的多维数据集。如何有效地从多维数据中提取或选择出有用的特征信息或规律,并将其分类识别已成为当今信息科学与技术所面临的基本问题。特征选择是指从原始特征集中选择使某种评估标准最优的特征子集,以使在该最优特征子集上所构建的分类或回归模型达到与特征选择前近似甚至更好的预测精度。RS[13](模拟退火算法)是一种用于特征选择、特征提取、特征减少和数据中决策规则提取的数学方法，特别是在数据不确定和不完整的情况下[14,15]。本文在粗糙集 RS 基础上提出特征最大依赖度算法（MDF-RS）算法进行特征选择，最后利用似然比检验[16]。

#### 2.2.1 RS 粗糙集理论

RS 是一种有效的数据处理方法，具有较强的分类能力。从而可以保持知识（即特征）分类不变的基础上对其进行简约。在文献[17]中，一个知识系统被定义为 $S=(U,A,V,f)$。其中：$U$ 是一个非空对象集；$A$ 是非空特征集；$V=U_{a\in A}V_a$，$V_a$ 是特征 $a$ 的值域；$f$：$U\times A\rightarrow V$ 是一个知识函数，即每一个 $(u,a)\in U\times A$ 时都有 $f(u,a)\in V_a$，即知识函数 $f$ 指定 $U$ 中每个对象 $u$ 的特征值。

**定义 1** 令 $S=(U,A,V,f)$ 是一个知识系统，$B$ 是 $A$ 的任意子

集，对于 $x,y\in U$，当且对每一个特征 $a\in B, f(x,a)=f(y,a)$ 则称 $x,y$ 关于 $B$ 是不可辨识关系，记为 $IND(B)$。很显然，$A$ 的每一个子集可以导出一个唯一的不可辨识关系，又称等价关系，而等价关系可以导出一个唯一的聚类，由 $IND(B)$ 导出的 $U$ 的聚类记为 $U/B$，聚类 $U/B$ 中包含 $x\in U$ 的等价类，记为 $[x]_B$。

**定义 2** 在知识系统 $S=(U,A,V,f)$ 中，$B$ 是 $A$ 的任意子集，$X$ 是 $U$ 的任意子集，把 $X$ 的 $B$ 下近似记为 $\underline{B}(X)$，$X$ 的 $B$ 上近似记为

$$\underline{B}(X)=\{X\in U|[x]_B\subseteq X\}, \overline{B}(X)=\{X\in U|[x]_B\cap X=\phi\} \tag{2}$$

可以看出 $\overline{B}(X)$ 可以用 $X$ 的补集 $(-X)$ 的下近似表示如式(3)所示，$U$ 的任意子集 $X$ 关于 $B$ 的近似精确度表示如式（4）所示。

$$\overline{B}(X)=U-B(-X) \tag{3}$$

$$\alpha_B(X)=|\underline{B}(X)/|\overline{B}(X)| \tag{4}$$

这里 $|X|$ 是集合 $X$ 的基数，即集合 $X$ 的元素个数。对于空集定义 $\alpha_B(\phi)=1$，很明显 $0\le\alpha_B(X)\le 1$。如果 $X$ 是 $U$ 的某些等价类的并集，那么 $\alpha_B(X)=1$，这时说集合 $X$ 关于 $B$ 是精确的。相反，如果 $X$ 不是 $U$ 的某些等价类的并集时，$\alpha_B(X)<1$，这时说集合 $X$ 关于 $B$ 不是精确的。这就意味着近似精确度 $\alpha_B(X)$ 越高，子集 $X\subseteq U$ 就越精确。

### 2.2.2 特征依赖度

在粗糙集理论中，可以这样理解特征重要度：一个知识系统 $S=(U,A)$ 中，$X\in A$ 是一个特征子集。如果 $x\subseteq A$，在 $X$ 中增加 $x$ 之后，知识系统提高了对对象的分辨能力，这种能力的提高程度就是特征重要度。提高程度越大，则 $x$ 对 $X$ 就越重要。通过特征依赖度可以发现，特征之间的内在联系重要特征之间的依赖度很小，重要特征与次重要特征之间的依赖度却较强，不重要特征与重要和次重要特征之间的依赖度很小。由此可以通过特征依赖度去除那些对分类不重要的特征或者提取出重要特征。

**定义 3** 在知识系统 $S=(U,A,V,f)$ 中，集合 $D$ 和 $C$ 是特征集合 $A$ 的任意子集，如果 $D$ 中的每一个值都可以精确到与 $C$ 的一个值关联，则称 $D$ 对 $C$ 是函数依赖的，记为 $C\Rightarrow D$。如公式(5)，令 $k$ 为依赖度，$D$ 以 $k$ 度依赖于 $C$，记为 $C\Rightarrow_k D$。如果 $k=1$，则 $D$ 完全依赖于 $C$；$k<1$，则 $D$ 部分依赖于 $C$

$$k=\sum_{X\in U/D}|\underline{C}(X)|/|U| \tag{5}$$

系数 $k$ 描述了通过特征 $C$ 能够将 $U$ 中的元素正确分类到划分 $U/D$ 的块中的比率。因此，当 $k=1$，$U$ 的全部或部分元素能够被划分到 $U/D$ 的等价类中。$k=0$ 时，$U$ 中没有元素能通过特征 $C$ 划分到 $U/D$ 的等价类中。也就是说特征间的依赖度越大对划分的决策影响越大。

### 2.2.3 特征最大依赖度算法（MDF-RS）

由于特征依赖度越大，特征越重要，对划分决策的影响就越大，因此，特征最大的依赖度算法的目标就是选出依赖度最大的特征作为分类的特征属性。具体算法步骤如下：

a) 对每个特征利用不可辨识关系计算等价类；

b) 用式（5）计算特征 $a_i\,(i\ne j)$ 的特征依赖度；

c) 选择每个特征的最大依赖度；

d) 根据特征属性的依赖度选取依赖度最大的属性作为分类特征属性。

最大依赖度选择举例：假设有 4 个属性 A,B,C,D，它们之间的依赖度如表 1 所示。

表 1 最大依赖度选择表

| 属性（依赖于） | 依赖度 k | 最大依赖度 |
|---|---|---|
| A | B 0.2　C 0.2 D 1 | 10.2 |
| B | A 0.4　C 0.2 D 1 | 10.4 |
| C | A 0.4　B 0.2 D 0.6 | 0.6 |
| D | A 0.4　C 0.2 B 0.2 | 0.4 |

比较表 1 中全部依赖度 k，可以发现最大的 k 是 1 出现在属性 A 和 B 上，然后再比较属性 A,B 的其它依赖度，发现最大的 k 等于 0.4 时出现在属性 B（A=0.4）上，由此选择 B 是分类特征属性。

### 2.2.4 基于 MDF-RS 的特征选择

本文将上述提出的 MDF-RS 算法进行 COPD 多维特征选取，选取过程如下：

a) 特征聚类。聚类的目的是将功能相近的特征聚在一起。为了提取低冗余度的特征，利用 K-均值聚类算法对最原始的数据特征进行聚类分析。其中，欧氏距离来度量两点之间的距离，并使用误差平方和(SSE)作为聚类的目标函数寻求最小的 SSE，如公式 6、7 所示，$k$ 表示 $k$ 个聚类中心，$c_i$ 表示第几个中心，$dist$ 表示的是欧几里德距离。

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{6}$$

$$SSE=\sum_{i=1}^{k}\sum_{x\in c_i}dist(c_i,x)^2 \tag{7}$$

b) 主特征选取。特征聚类后，每组类别中包含的特征功能是相似的，因此选取主要特征来表示这个类别并汇合这些主要特征构成的特征组。COPD 特征选择方法描述如下。

| Algorithm1:Feature selection of COPD |
|---|
| 1.Input:Sample set |
| 2.Output Feature G |
| 3.Clustering，get $\{A_1,A_2..A_k\}$; |
| 4. $G\ne\phi$; |
| 5.FOR( $i=1, i\le k, i++$ ) |
| 6.{ a:Calculate　$g\in A_i$ Sample equivalent class; |
| 7.　b:Calculate the degree of feature dependence; |
| 8.　c:Compare　$k_i$ , $g_i$ as $A_i$ category main features; |

```
9.  d: G = G + g_i ;
10.}return G
```

## 2.3 直接搜索模拟退火算法 DSA

DSA[18](直接搜索模拟退火算法)，是对 SA[19](模拟退火算法)的改进,该算法在两个方面区别于 SA。首先在 SA 中，算法只维持一个当前最优点，而在 DSA 过程中，算法维持一个工作点集合。所以在 SA 中，算法只在一个点附近搜索，这使得 SA 可能会陷入局部收敛，而在 DSA 中，算法在一组工作点集合附近搜索，从而能有效地跳出局部最优点。改进后的 DSA 算法如算法 2 所示。

Algorithm2:A direct search variant of the simulated annealing algorithm

```
1.G=G0,p=P(s);//Initial state, precision
2. Gbest=G,pbest=p;
3.k=0；kmax=Constant Value;
4.MaxScore=A constant Value;    //evaluation count
5.while（k<kmax & p<=MaxScore）
6.{//While time left & not good enough
7.    Gnew=Neighbor(G);
8.    pnew=P(Gnew);
9.    if exp (pnew - p) > Random ()
10.   {
11.        G=Gnew;
12.        p=pnew;  //Yes, change state.
13.   }
14    if pnew>pbest
15    {
16.         Gbest=Gnew;
//Save'new neighbouring' 'best found
17.         pbest=pnew;
18.         k=k+1;
19.     }
20.}return Gbest, pbest    //Return the best solution found.
```

初始化 DSA 的参数，然后随机初始化 SVM 的参数(C,γ)。首先为它们选择邻居，并尝试用 DSA 搜索来调整这个邻居，通过交叉验证技术来比较这些不同的（C,γ）为了不断优化参数（C,γ）。其次，为了进一步调整内核函数参数，我们在最佳局部（C,γ）周围构建一个虚拟窗口，直到该参数为我们所接受范围内，当调整 C 和 γ 的参数值使得准确率等指标不断提高并趋于稳定时停止调参。最后，使用最优的（C,γ）参数 DSA-SVM 建立模型并测试数据集。本文参数（C,γ）的间隔区间设置为（$2^{-5}$,$2^{-15}$），($2^{-15}$,$2^{-5}$),对于所有可能参数组合（C,γ）用交叉验证计算。随后，解释一下本文在 DSA 直接搜索模拟退火算法使用交叉验证算法。k 折交叉验证的算法来优化参数，具体步骤如下所示：

a）随机将样本集 S 划分成 k 个不相交的子集，每个子集中样本数量为 m/k 个，这些子集分别记作 $S_1,S_2..S_k$；

b) 对于每个模型，进行如下操作：for j=1 to k, $S_1...\cup S_{j-1}...\cup S_{j+1}...\cup S_k$ 作为训练集，训练模型 $C_i = A_\beta^j(S_j \setminus S)$

c)计算每个模型的平均泛化误差，根据式（8），选择泛化误差最小的模型 $C_i$。K 折交叉验证方法，每次留作验证的为总样本量的 1/k.

$$\zeta_{MES} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{8}$$

通过交叉验证得到的每组($c,v$)组合，公式如式（9）所示。

$$CVS = \frac{\#\ predicted \quad records}{\#total \quad records} \tag{9}$$

## 2.4 建立模型

本文构建的分类器模型及方法的流程图如图 1 所示，首先模型输出(C, γ)的最优值,然后构建分类器。在获得最好的数据对(C, γ)之后,构建双向耦合(PWC)概率估计的学习分类器。双向耦合是一个受欢迎的多层次分类方法，它将对每个类的所有比较组合了起来。PWC 构造了 $r_{ij}=k(k-1)/2$, 1≤i≤k, 1≤j≤≤i 的分类器。这个分类决策是由聚合分类器的输出做出的。

二元分类器用于估计成对类的概率 $\mu_{ij}^* = p(Y_0 = i|Y_0 = i or Y_0 = j, x_0)$， $r_{ij}$ 对 $\mu_{ij}^*$ 的估计可以通过训练训练集的第 $i$ 个和第 $j$ 个类得到。为了计算这个概率，我们用了杜[20]的方法。

然后，使用所有的 $r_{ij}$ 来达到目标，即估计 $p_i^* = (Y_0 = i)|x_0$，i=1…K。因此，在测试阶段，每个分类器都可以估计分类结果的概率,如式（10）所示。

$$d_{ij} = \{(x_n, y_n)|y_n = i\ or \quad y_n = j, 1 \le n \le N\} \tag{10}$$

# 3 实验结果

## 3.1 COPD 数据集

COPD 数据集是从合作伙伴医疗系统的电子医疗记录中提取的，并且筛选出对患者观察至少 5 年的数据作为我们的实验数据集。该实验的目的是通过对患者进行各种医学检测的结果及症状表现来预测 COPD 疾病是否存在.数据集含有 1200 个样本，属于两个不同类别，共有 750 名 COPD 患者（62.5%）和 450 名（37.5%）不是 COPD 患者但与 COPD 患者有相似症状，我们从实验样本的电子病历中提取出原始的 26 个特征。特征的描述如表 2 所示。

## 3.2 COPD 特征选择结果

原始数据集中对高维特征提取是模型精确度等各项指标的关键一步，所以，对原始特征选择对于分类模型具有重要意义。但在传统学习方法中无法提取最优的特征组合，因此，本文提出了 MDF-RS 算法并通过该算法获得的特征子集。

a)特征聚类。在前面我们介绍使用 K-均值聚类算法，根据原始数据的特点，本实验初始化 k=7，把原始数据聚为 7 堆，结果如图 2 所示。

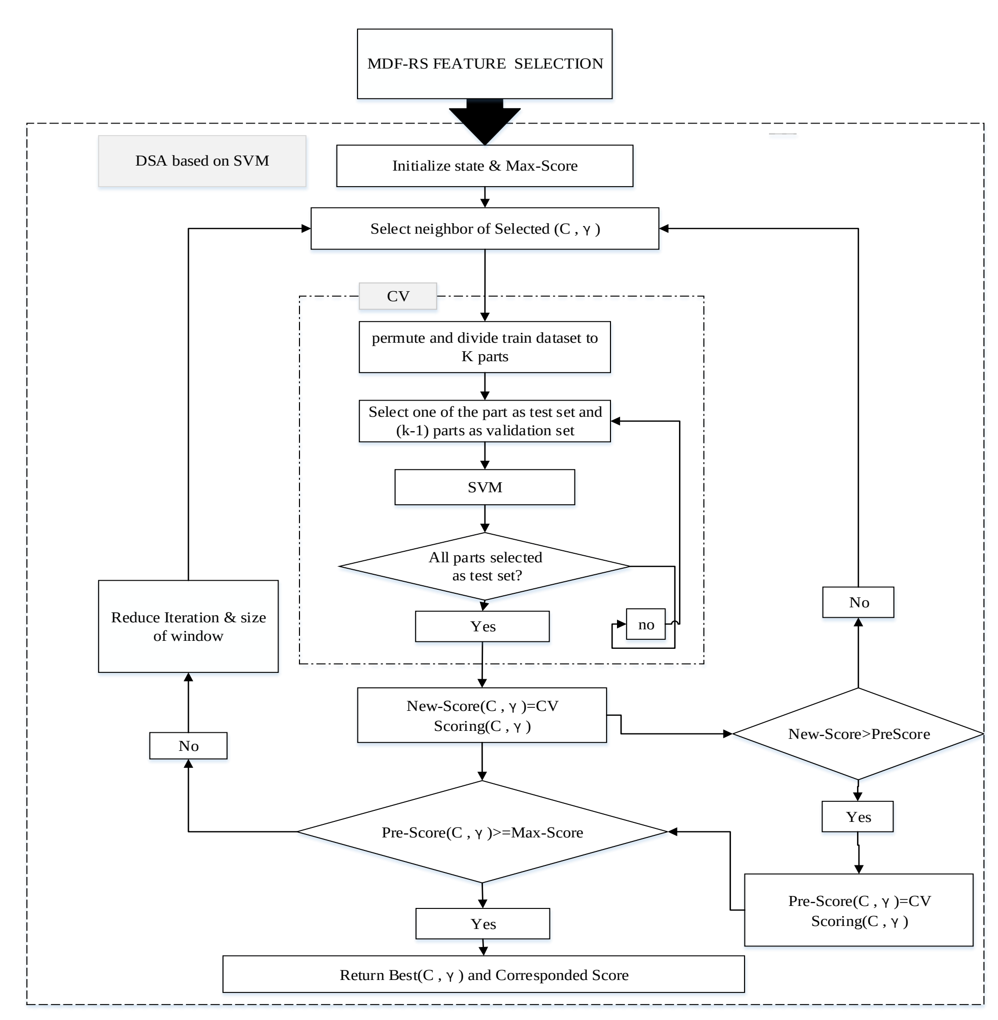



图 1　DSA-SVM 分类模型

表 2　COPD 特征表

| 特征 | 特征值 | 特征 | 特征值 |
|---|---|---|---|
| F0:Sex | Male,female | F13:咽干 | 0,1 |
| F1:FEV1/FVC | 0~1 | F14:咳嗽 | 0,1,2,3,4,5 |
| F2:劳动 | 0,1,2,3,4,5 | F15:畏热 | 0,1 |
| F3:流涕 | 0,1 | F16:胸痛 | 0,1 |
| F4::Age | 0,1,2,3,4,5 | F17:胸闷 | 0,1,2,3,4,5 |
| F5:乏力 | 0,1 | F18:心慌 | 0,1 |
| F6:自汗 | 0,1 | F19:信心 | 0,1,2,3,4,5 |
| F7:咳痰 | 0,1,2,3,4,5 | F20:精力 | 0,1,2,3,4,5 |
| F8:睡眠 | 0,1,2,3,4,5 | F21:发热 | 0,1 |
| F9:抽烟 | 0,1 | F22:咽痒 | 0,1 |
| F10:体重 | 0,1,2,3,4,5 | F23:浮肿 | 0,1 |
| F11:便秘 | 0,1 | F24:舌苔 | 0,1 |
| F12:mMRC | 0,1,2,3,4，5 | F25:紫绀 | 0,1 |

b)主特征选取。根据聚类后的 7 堆特征，使用 MDF-RS 算法从中选取主要的特征作为特征子集。从表 3 可以看出，特征组合是由 9 到 19 维数的特征子集组成，通过 MDF-RS 算获得了 14 个的子特征组合（R1-R14）。特征权重归一化后，特征按权重排序如图 3 所示.提取的最优特征子集组合将作为 DSA-SVM 模型的输入，最后用自然比检验，计算结果如表 4 所示。其中似然比检验统计量的公式如式（11）所示。

$$G=-2\left[\ln\left(L_{m-1}\right)-\ln\left(L_m\right)\right] \tag{11}$$

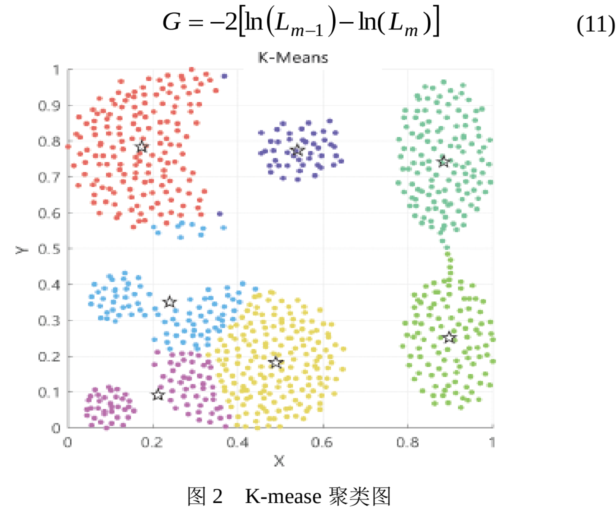



图 2　K-mease 聚类图

表 3　特征选择

| R | size | feature |
|---|---|---|
| R1 | 9 | $F_0,F_2,F_3,F_4,F_5,F_8,F_{10},F_{14},F_{18}$ |
| R2 | 9 | $F_0,F_1,F_3,F_6,F_9,F_{12},F_{17},F_{21},F_{23}$ |
| R3 | 12 | $F_0,F_2,F_4,F_5,F_8,F_{11},F_{13},F_{16},F_{18},F_{19},F_{21},F_{23}$ |
| R4 | 13 | $F_1,F_2,F_4,F_6,F_7,F_8,F_{10},F_{11},F_{13},F_{14},F_{16},F_{17},F_{18}$ |
| R5 | 13 | $F_0,F_2,F_3,F_4,F_6,F_7,F_{10},F_{11},F_{14},F_{15},F_{18},F_{21},F_{24}$ |
| R6 | 14 | $F_0,F_2,F_3,F_5,F_7,F_8,F_{10},F_{11},F_{14},F_{16},F_{18},F_{21},F_{22},F_{25}$ |
| R7 | 15 | $F_0,F_1,F_2,F_4,F_5,F_6,F_8,F_{10},F_{12},F_{13},F_{16},F_{18},F_{20},F_{21},F_{24}$ |
| R8 | 16 | $F_0,F_1,F_2,F_3,F_6,F_8,F_9,F_{10},F_{11},F_{12},F_{15},F_{16},F_{18},F_{19},F_{21},F_{24}$ |

| | | |
|---|---|---|
| R9 | 17 | $F_0,F_1,F_2,F_3,F_5,F_6,F_7,F_8,F_{11},F_{13},F_{15},F_{17},F_{18},F_{19},F_{23},F_{24},F_{25}$ |
| R10 | 17 | $F_0,F_1,F_2,F_3,F_5,F_6,F_7,F_8,F_{11},F_{12},F_{16},F_{17},F_{18},F_{19},F_{23},F_{24},F_{25}$ |
| R11 | 18 | $F_0,F_1,F_2,F_3,F_5,F_6,F_7,F_8,F_{11},F_{14},F_{16},F_{17},F_{18},F_{19},F_{22},F_{23},F_{24},F_{25}$ |
| R12 | 19 | $F_0,F_1,F_2,F_3,F_4,F_5,F_6,F_7,F_8,F_{11},F_{14},F_{16},F_{17},F_{18},F_{19},F_{22},F_{23},F_{24},F_{25}$ |
| R13 | 19 | $F_0,F_1,F_2,F_3,F_4,F_5,F_6,F_7,F_8,F_9,F_{10},F_{11},F_{12},F_{13},F_{14},F_{15},F_{16},F_{17},F_{18}$ |
| R14 | 19 | $F_1,F_2,F_3,F_4,F_5,F_6,F_7,F_9,F_{10},F_{11},F_{14},F_{15},F_{17},F_{18},F_{19},F_{21},F_{23},F_{24},F_5$ |

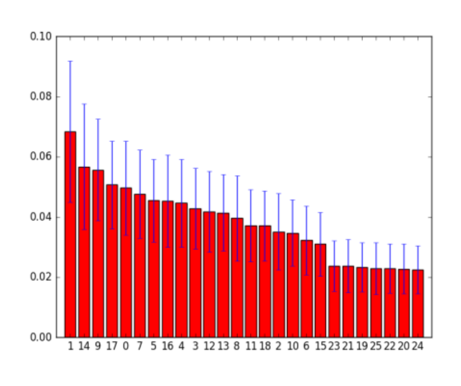



图 3　特征选择图

表 4　似然比测试表

| | |
|---|---|
| 特征 | $F_0,F_1,F_2,F_3,F_4,F_5,F_6,F_7,F_8,F_9,F_{10},F_{11},F_{12},F_{13},F_{14},F_{15},F_{16},F_{17},F_{18}$ |
| G | 8.2,7,6.5,6,6.3,5.5,5.3,5,4.9,4.7,4,4.6,4.8,4.9,4.8,4,3.8,3.9,3.6 |

从表 4 结果可以看出，在 19 个检验统计量都大于 $\chi^2_{1,0.05}=3.84$，且 $p \prec 0.05$ 说明有统计意义，这与通过 MDF-RS 算法特征选择出的特征组合 R13 一致。因此，通过结果得出在其中一个变量在其他 18 个变量不变的情况下影响显著，所以选取的这 19 个多维特征对慢阻肺诊断非常有意义。

### 3.3　实验结果 DSA-SVM

本文利用 MDF-RS 算法进行特征选择后，并通过 DSA-SVM 对数据集进行分类。为了提高模型的准确率等各项指标，参数 $C$ 和 $\gamma$ 组合搭配是非常重要的，因此，利用直接搜索模拟退火算法对 SVM 参数 $C$ 和 $\gamma$ 组合进行优化，本文在局部参数内建立一个虚拟窗口，并设置参数范围阈值直到参数为所接受范围趋于稳定，最后用交叉验证方法找到参数 $C$ 和 $\gamma$ 最优组合。我们对本实验的 $C$ 和 $\gamma$ 的参数组合及对应准确率用三维图表示，如图 4 所示。其中，图 4 用方框围起的点就是参数最优组合及其准确率，$C$ 和 $\gamma$ 分别为（14.5，0.352），准确率达到 94.6%，而不同特征组合 R 的分类准确率如表 5 所示。

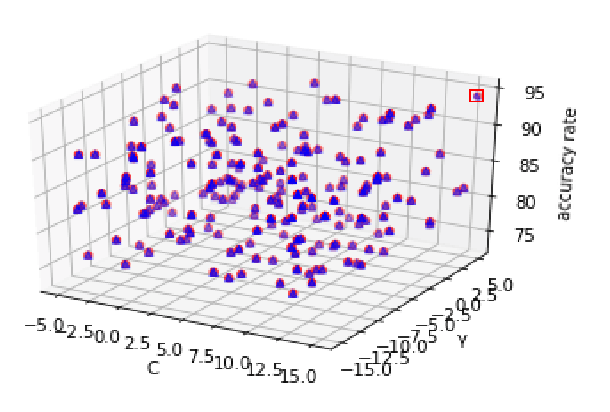



图 4　$C$、$\gamma$ 及准确率三维表示图

表 5　分类指标精度

| R | 准确率 | 特异性 | 灵敏度 |
|---|---|---|---|
| R1 | 89.25% | 94.21% | 97.42% |
| R2 | 86.13% | 94.14% | 96.33% |
| R3 | 88.62% | 93.25% | 97.15% |
| R4 | 87.21% | 94.56% | 98.22% |
| R5 | 90.38% | 95.34% | 97.43% |
| R6 | 91.53% | 93.48% | 96.84% |
| R7 | 90.34% | 95.82% | 94.16% |
| R8 | 94.22% | 96.31% | 98.36% |
| R9 | 94.17% | 96.62% | 99.21% |
| R10 | 93.84% | 95.81% | 98.34% |
| R11 | 92.2% | 96.95% | 98.89% |
| R12 | 94.13% | 94.83% | 97.37% |
| R13 | 94.6% | 96.2% | 99.83% |
| R14 | 94.52% | 94.97% | 97.34% |

### 3.4　实验比较

在文献中，有大量研究者用单一的和混合方法来诊断慢阻肺疾病，但在处理数据集缺失值及模型参数方面存在着不足。文本通过 MDF-RS 特征提取算法和 DSA-SVM 分类模型对慢阻肺诊断取得了良好的效果。

首先，在本节中，所提出的方法与先前的机器学习模型比较来进行比较。本文的 DSA-SVM 算法在准确率、召回率、F1 值三个指标都取得了良好的效果，比较结果如表 6 所示。

表 6　方法比较

| 方法 | 准确率 | 召回率 | F1 |
|---|---|---|---|
| Logistic | 90.3% | 85.65% | 89.2% |
| Decisiontree | 92.26% | 83.7% | 87.43% |
| XGBoost | 93.76% | 87.8% | 90.4% |
| 随机森林 | 93.68% | 92.4% | 89.7% |
| Svm | 93.7% | 91.32% | 91.21% |
| DSA-svm | 94.6% | 93.2% | 92.9% |

其次，在本节中除了与不同模型之间的比较，还与 H 文献[3，5]进行了比较。Himes 使用了贝叶斯网络模型预测 COPD 患者,准确率达到 83.3%，而郭等人使用了 ARIMA 模型准确率达到 90.2，本文相比较 Himes 和郭的准确率、F1 值有所提高，同时 AUC 值达到了 0.94，这说明本文方法取得了有效结果。准确率、F1 值、ROC 对比图如图 5~7 所示。

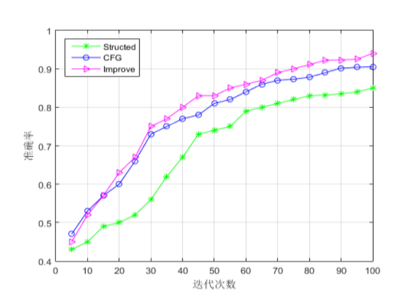



图 5　准确率比较图

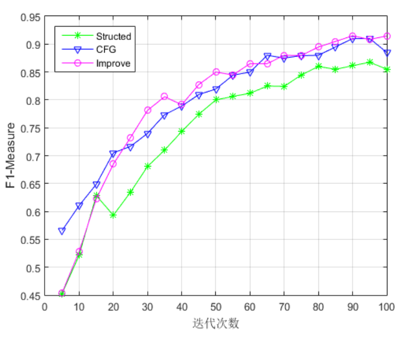


图 6 $F_1$ 指数比较图

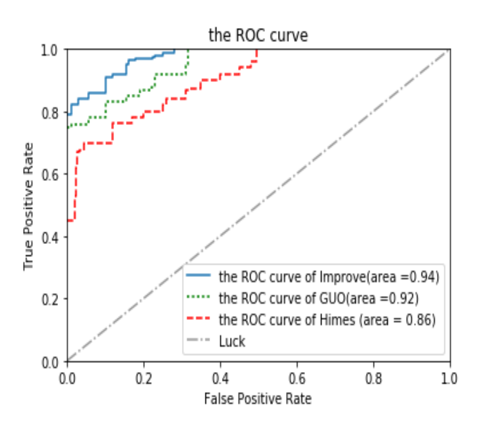


图 7 ROC 比较图

## 4 结束语

本文提出了用于慢阻肺疾病诊断预测的 DSA-SVM 模型及 MDF-RS 多维特征选择算法，并通过各种指标进行比较，从这些结果可以看出，用于慢阻肺疾病的 DSA-SVM 诊断算法获得了较好的结果。因此，所提出的 DSA-SVM 诊断算法对于医生对患者做出最终决定是非常有帮助的，它可以辅助医生对慢阻肺疾病进行诊断从而减少误诊率。在未来的慢阻肺疾病诊断研究中，将使用不同的特征提取和其他学习方法来提高诊断系统的准确性。

## 参考文献：

[1] Lópezcampos J L, Tan W, Soriano J B. Global burden of COPD [J]. Journal of Applied Mathematics, 2016, 21 (1): 14.

[2] Celik M U, Sharma G, Tekalp A M. Lossless watermarking for image authentication: a new framework and an implementation [J]. IEEE Trans on Image Processing, 2006, 15 (4): 1042-1049.

[3] Himes B E, Dai Y, Kohane I S, *et al.* Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records [J]. Journal of the American Medical Informatics Association, 2009, 16 (3): 371-379.

[4] Hoogendoorn M, Feenstra T L, Boland M, *et al.* Prediction models for exacerbations in different COPD patient populations: comparing results of five large data sources [J]. International Journal of Chronic Obstructive Pulmonary Disease, 2017, 12 (5): 3183-3194.

[5] 郭慧敏，杜军，黄路非. 基于 R 语言 ARIMA 模型在慢阻肺急性加重患者发病预测中的应用 [J]. 中国卫生统计，2017, 34 (2): 288-289. (Guo Huimin, Du Jun, Huang Lufei. Application of ARIMA model based on R language in predicting incidence of patients with acute exacerbation of chronic obstructive pulmonary disease [J]. Chinese Health Statistics, 2017, 34 (2): 288-289. )

[6] Steyerberg E W. Clinical Prediction Models [M]. Springer US, 2010.

[7] Moons K G M, Royston P, Vergouwe Y, *et al.* Prognosis and prognostic research: what, why, and how? [J]. Bmj, 2009, 338 (7706): b375-b383.

[8] Steyerberg E W, Moons K G M, Windt D A V D, *et al.* Prognosis research strategy (progress) 3: prognostic model research [J]. PLoS Medicine, 2013, 10 (2): e1001381-e1001389.

[9] Moons K G M, Kengne A P, Grobbee D E, *et al.* Risk prediction models: II. External validation, model updating, and impact assessment [J]. Heart, 2012, 98 (9): 691-699.

[10] Bleeker S E, Moll H A, Steyerberg E W, *et al.* External validation is necessary in prediction research: a clinical example [J]. Journal of Clinical Epidemiology, 2003, 56 (9): 826-832.

[11] Mega J L, Braunwald E, Wiviott S D, *et al.* Rivaroxaban in patients with a recent acute coronary syndrome [J]. New England Journal of Medicine, 2012, 366 (1): 9-18.

[12] Cheung N. Machine learning techniques for medical analysis [J]. Journal of Clinical Epidemiology, 2017, 26 (4): 126-132.

[13] Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease [J]. Applied Soft Computing Journal, 2013, 13 (8): 3429-3438.

[14] Kaneiwa K. A rough set approach to multiple dataset analysis [M]. Elsevier Science Publishers B. V. 2011.

[15] Chen Y, Miao D, Wang R. A rough set approach to feature selection based on ant colony optimization [J]. Pattern Recognition Letters, 2010, 31 (3): 226-233.

[16] Shafiq M, Atif M, Viertl R. Generalized likelihood ratio test and cox's f-test based on fuzzy lifetime data [M]. John Wiley & Sons, Inc. 2017.

[17] Thangavel K, Pethalakshmi A. Dimensionality reduction based on rough set theory: a review [J]. Applied Soft Computing, 2009, 9 (1): 1-12.

[18] Ali M M, Törn A, Viitanen S. A direct search variant of the simulated annealing algorithm for optimization involving continuous variables [J]. Computers & Operations Research, 2002, 29 (1): 87-102.

[19] Sartakhti J S, Zangooei M H, Mozafari K. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA) [J]. Computer Methods & Programs in Biomedicine, 2012, 108 (2): 570-579.

[20] 杜占龙，李小民，席雷平，等. 多分类概率极限学习机及其在剩余使用寿命预测中的应用 [J]. 系统工程与电子技术，2015, 37 (12): 2777-2784. (DU Zhanlong, LI Xiaomin, XI Leiping. Multi-class probabilistic extreme learning machine and its application in remaining useful life prediction [J]. Systems Engineering and Electronics, 2015, 37 (12): 2777-2784. )